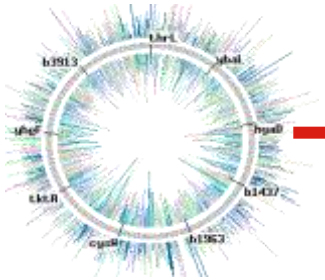
Standards in Basic Genomics Research – Lessons Learned

Alvis Brazma
European Bioinformatics Institute, EMBL
Wellcome Trust Genome Campus, Cambridge, UK

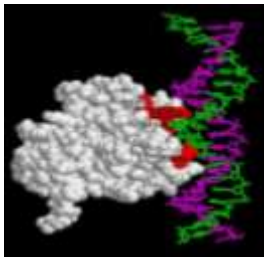
Biomedical Informatics without borders
NCI/NCRI Workshop, Bethesda
June 21-22, 2010

EMBL-EBI

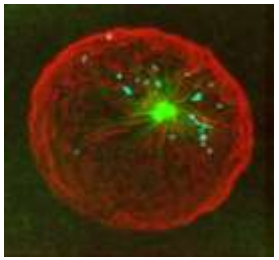
Biomedical research requires data integration



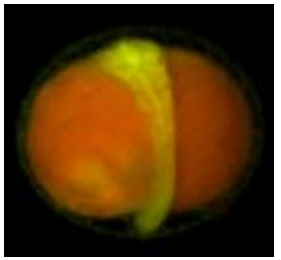
Genome



Protein



Cell

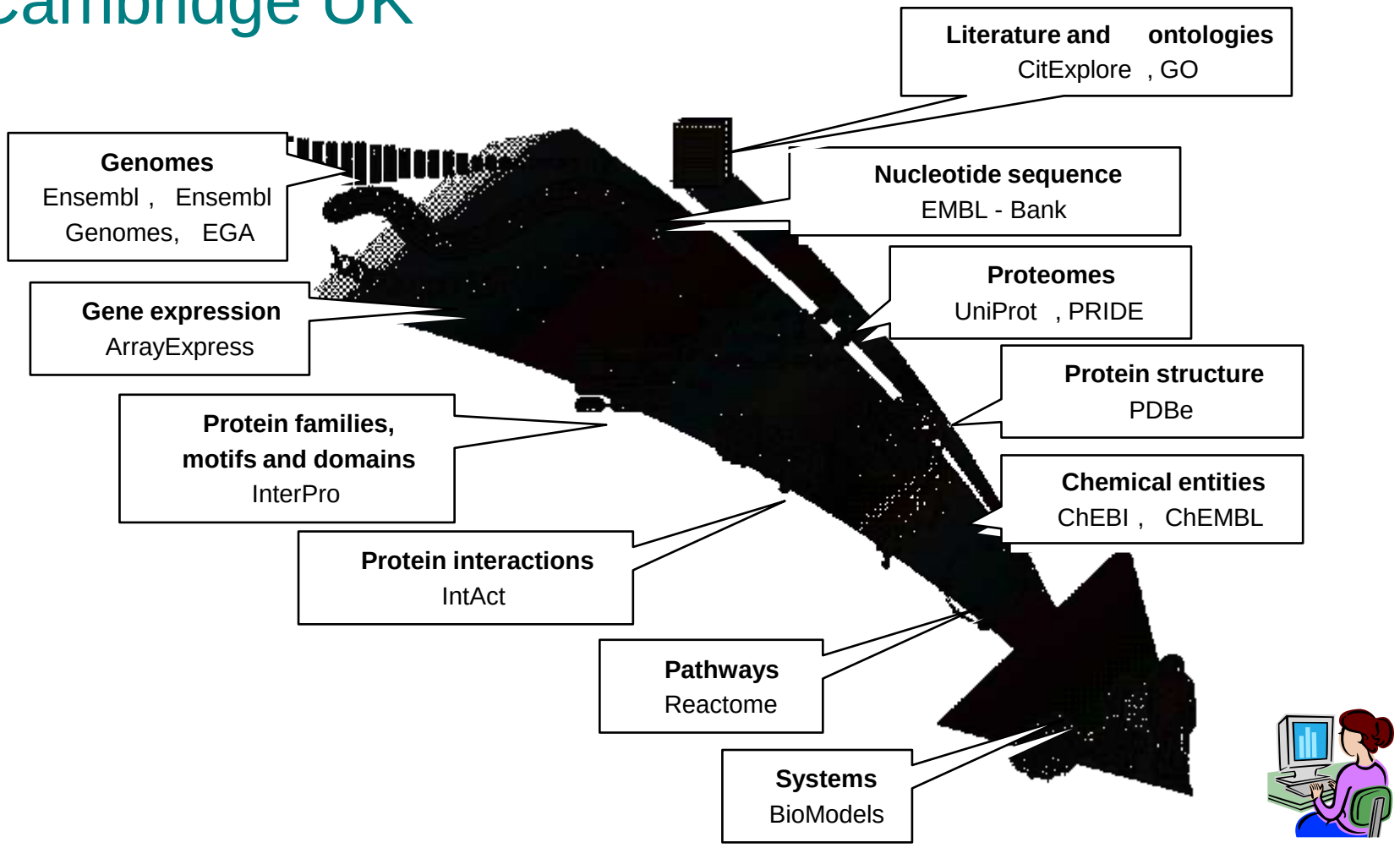


Embryo

Model organisms

Human health

European Bioinformatics Institute
Wellcome Trust Genome Campus
Cambridge UK



EMBL-EBI: Most important data collections

Genomes & Genes

- **Ensembl:** Joint project with Sanger Institute - high-quality annotation of vertebrate genomes
- **Ensembl Genomes:** Environment for genome data from other taxons
- **1000 Genomes:** Catalogue of human variation from major World populations
- **EGA*:** European Genotype Archive* – genotype, phenotype and sequences from individual subjects and controls
- **ENA:** European Nucleotide Archive – all DNA & RNA, nextgen reads and traces

Transcription

- **ArrayExpress:** Archive of transcriptomics and other functional genomics data
- **Expression Atlas:** Differentially expressed genes in tissues, cells, disease states & treatments

Protein

- **UniProt:** Archive of protein sequences and functional annotation
- **InterPro:** Integrated resource for protein families, motifs and domains
- **PRIDE:** Public data repository for proteomics data
- **PDBe:** Protein and other macromolecular structure and function

Small molecules

- **ChEBI:** Chemical entities of biological interest
- **ChEMBL:** Bioactive compounds, drugs and drug-like molecules, properties and activities

Processes

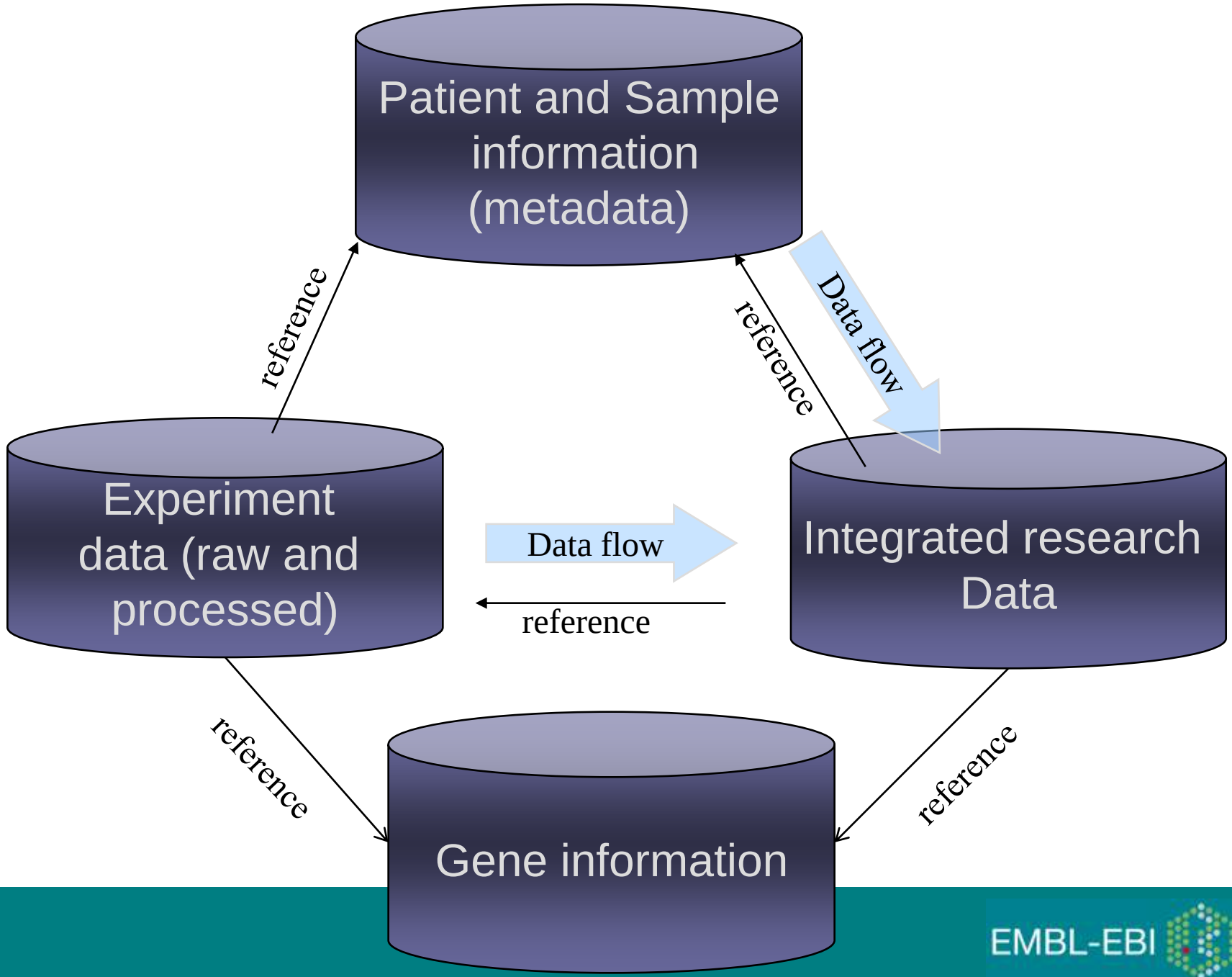
- **IntAct:** Public repository for molecular interaction data
- **Reactome:** Biochemical pathways and reactions in human biology
- **Biomodels:** Mathematical models of cellular processes

Ontologies

- **GO:** Gene Ontology, consistent descriptions of gene products

Scientific literature

- **CiteXplor:** Bibliographic query system


Patient and Sample information (metadata)
reference
reference
Data flow
Experiment data (raw and processed)
Data flow
reference
Integrated research Data
reference
reference
Gene information

SCIENTIFIC CONFERENCES
HINXTON CAMBRIDGE UK

wellcome trust

Bridging the Gap in Biomedical Genetics

27–29 October 2010

Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

BACKGROUND

Biomedical research is increasingly using new molecular technologies, such as genotyping, transcription profiling, or metabolite profiling to understand disease mechanisms, test drug efficacy, and develop new therapies. Patient samples are now routinely profiled by such technologies and the data are collected in biomedical research warehouses for analysis. The patient health records are typically managed in separate information systems. To achieve the full power of the new approaches for translational research, the molecular data need to be integrated with information from the patient records.

SCIENTIFIC ORGANISING COMMITTEE:

Janet Thornton European Bioinformatics Institute, UK
Alvis Brazma European Bioinformatics Institute, UK
David Ingram University College London, UK
Johan Van der Lei Erasmus University Medical Center, Netherlands

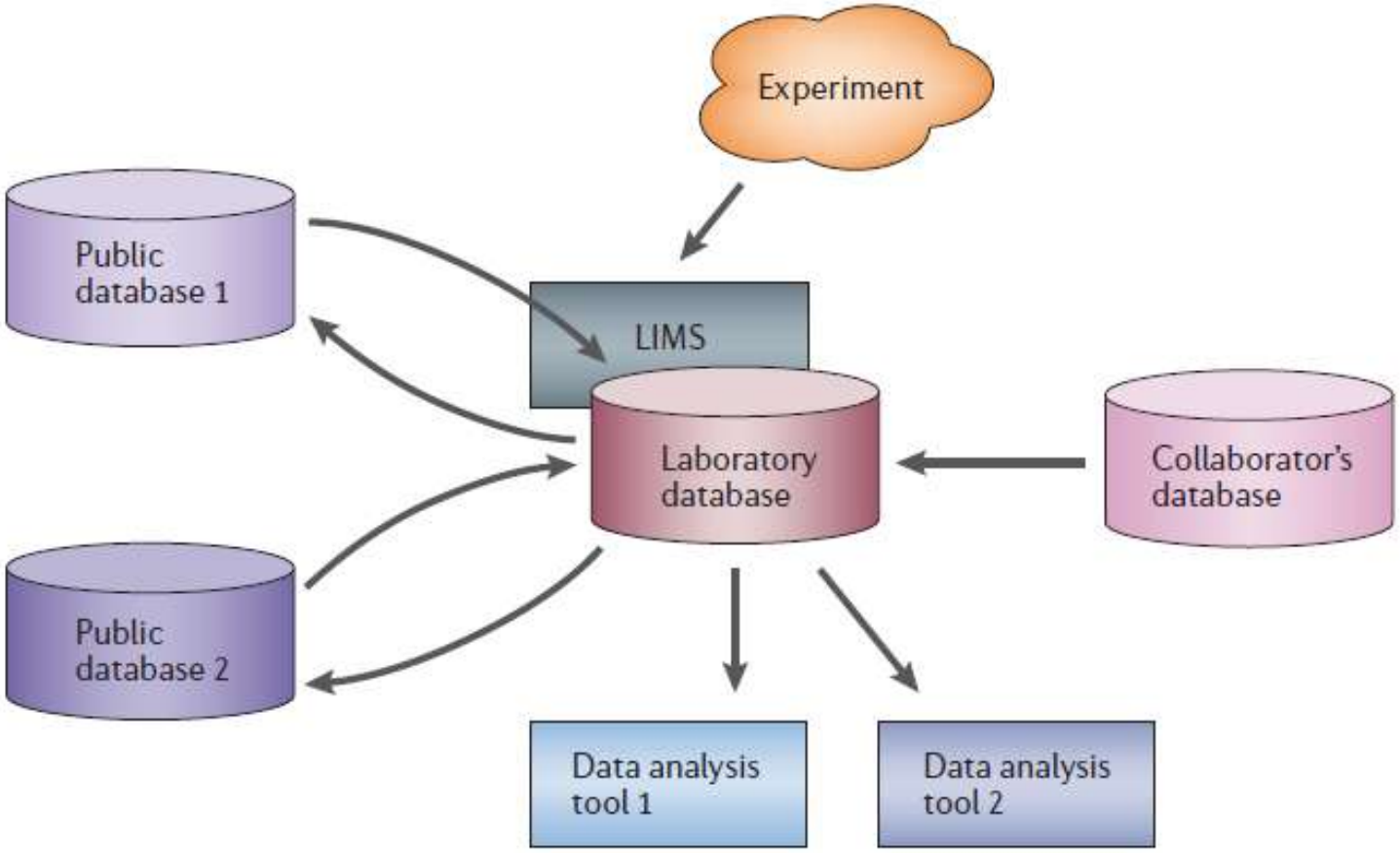
KEYNOTE SPEAKERS:

Norbert Graf University of Saarland, Germany
Isaac Kohane Harvard MIT, USA
Bernadette Modell University College London, UK
Gertjan van Ommen Leiden University Medical Center, Netherlands

Why do we need standards?

- To help to fit together many pieces to make the useful whole when we have different manufacturers and product providers
- Standards in IT – to exchange large volumes of data when many smaller transactions are needed and no single architecture can cover all applications
- Successful examples
  - Telecommunications
  - Banking

No single architecture can cover all biomedical information applications



Nature Genetics Reviews (2006)

Standards for systems biology

Alvis Brazma, Maria Krestyaninova and Ugis Sarkans

Abstract | High-throughput technologies are generating large amounts of complex data that have to be stored in databases, communicated to various data analysis tools and interpreted by scientists. Data representation and communication standards are needed to implement these steps efficiently. Here we give a classification of various standards related to systems biology and discuss various aspects of standardization in life sciences in general. Why are some standards more successful than others, what are the prerequisites for a standard to succeed and what are the possible pitfalls?

Domain
A field of study.

Barbarism is the absence of standards to which appeal can be made. José Ortega y Gasset

Historically, standards emerged from a need for a recommended practice in the manufacturing of products. The main purpose of standards is to help to fit together many pieces to make a useful whole. In information technology, standards are needed to exchange large volumes of information when many smaller transactions are needed. Examples of established standards include a Hypertext Transport Protocol (HTTP) and an Exensible Markup Language (XML) for structuring data.

In the life sciences, the advances in high-throughput

What is the aim of these standardization initiatives? How many of these standards are used and how many have contributed to advances in biology? The goal of this review is to help systems biologists to navigate the rapidly changing maze of data standards in life sciences, and to encourage biologists to provide feedback on these developments. Such feedback is necessary to ensure that the developed standards are as close as possible to addressing the real needs of biology.

It is important to distinguish between standards that specify how to actually do experiments and standards that specify how to describe experiments. Recommendations such as what standard reporters (probes) should be printed on microarrays or what quality control steps

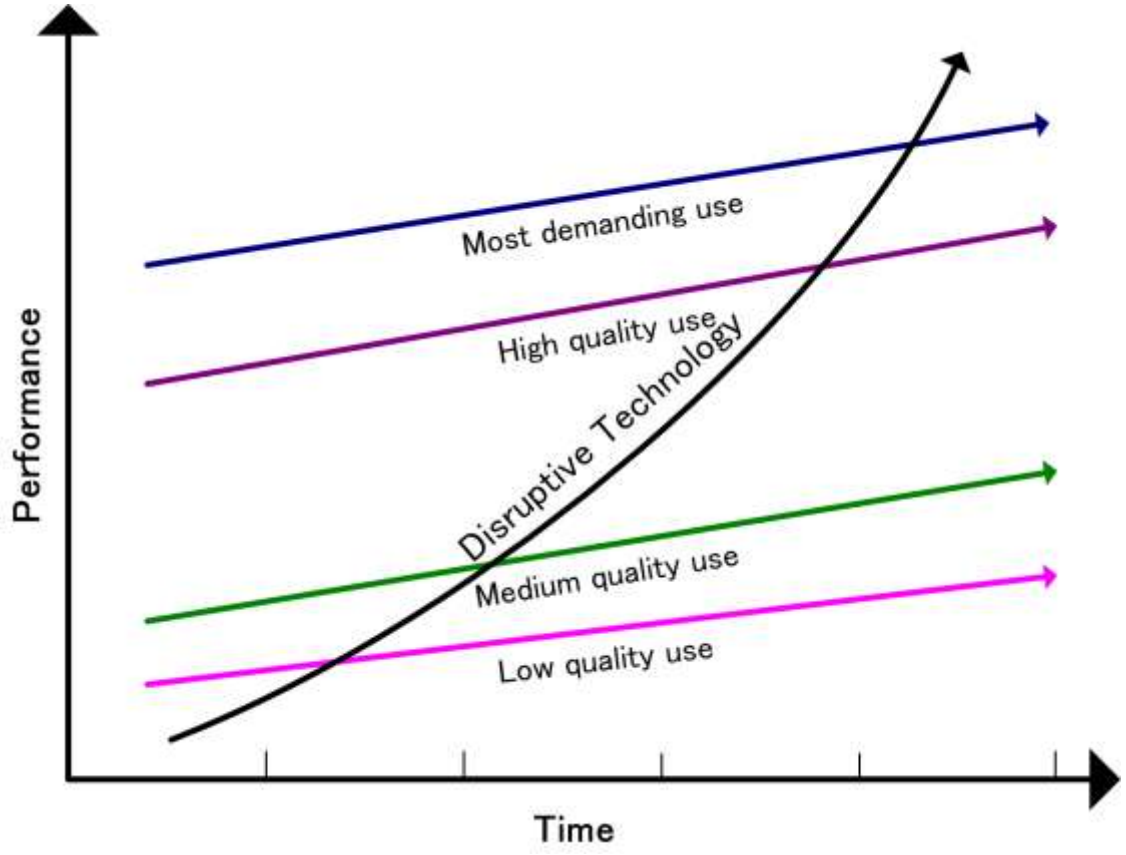
Four steps in developing a bioinformatics standard:

- Conceptual design and reporting requirements
- Model formalisation – ontology, object model, identifiers
- Data exchange format
- Supporting software

What's special about basic research?

- To be on the leading edge you need to do things differently from the accepted practice
- You have to apply the latest technologies, which keep appearing monthly
- The goal of basic research is to 'break out of the box'

Disruptive technologies.



"A technology becomes disruptive when the rate at which it improves exceeds the rate at which users can adapt to the new performance."

The Innovator's Dilemma. Clayton M. Christensen. Harvard Press. 1997

Disruptive technologies in biology

- Next-generation DNA sequencing
  - Data will be 1,000 <> 1,000,000 times cheaper to produce
  - Data production rates will be 1,000 <> 1,000,000 more by the end of the decade.
- Protein sequencing by Mass Spectrometry may also be disruptive
- There will probably be others
  - Macromolecular structure determination by Electron Microscopy
  - Imaging of various kinds
  - etc

Data in basic research changes somewhat faster than in banking

Franklin: Have you ever thought, Headmaster, that your standard might perhaps be a little bit out of date?

Headmaster: Of course they are out of date. Standards are always out of date. That is what makes them standards.

- *Forty Years On*, by Alan Bennett, 1968

So what do we do?

- Give up? Not an option
- Be modest – concentrate on areas with relatively established protocols and high throughput
- Try to make sure that use of standards give real benefits to the scientists (end users) quickly (be agile!)
- Successful examples in Genomics and Bioinformatics
  - Gene Ontology (GO)
  - Minimum Information About a Microarray Experiment (MIAME)
  - Systems Biology Markup Language (SBML)
  - R and Bioconductor (*Gene Set* objects)
  - BAM format for UHTS

How do we measure success?

- Citations?
- Usage/Uptake?
- Does this enable research that would not be possible otherwise (but how do you judge this?)

Gene Ontology (GO)

- The Gene Ontology, or GO, is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species

- Ashburner, CA Ball, JA Blake, D Botstein, H Butler … Gene Ontology: tool for the unification of biology - *Nature genetics*, 2000 - **Cited by 5869** (Google Scholar)

- Almost all high throughput data analysis papers use GO in some way, analysis of groups of hundreds of genes would not be possible without GO (eg. Gene set enrichment analysis)

Minimum information about a microarray experiment (MIAME)—toward standards for microarray data

Alvis Brazma[1], Pascal Hingamp[2], John Quackenbush[3], Gavin Sherlock[4], Paul Spellman[5], Chris Stoeckert[6], John Aach[7], Wilhelm Ansorge[8], Catherine A. Ball[4], Helen C. Causton[9], Terry Gaasterland[10], Patrick Glenisson[11], Frank C.P. Holstege[12], Irene F. Kim[4], Victor Markowitz[13], John C. Matese[4], Helen Parkinson[1], Alan Robinson[1], Ugis Sarkans[1], Steffen Schulze-Kremer[14], Jason Stewart[15], Ronald Taylor[16], Jaak Vilo[1] & Martin Vingron[17]

Microarray analysis has become a widely used tool for the generation of gene expression data on a genomic scale. Although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. Here we present a proposal, the Minimum Information About a Microarray Experiment (MIAME), that describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. The ultimate goal of this work is to establish a standard for recording and reporting microarray-based gene expression data, which will in turn facilitate the establishment of databases and public repositories and enable the development of data analysis tools. With respect to MIAME, we concentrate on defining the content and structure of the necessary information rather than the technical format for capturing it.

Introduction

After genome sequencing, DNA microarray analysis[1] has become the most widely used source of genome-scale data in the life sci-

cult, because at present, microarrays do not measure gene expression levels in any objective units. In fact, most measurements report only relative changes in gene expression, using a reference which is

NATURE VOL 403 17 FEBRUARY 2000

commentary

One-stop shop for microarray data

Is a universal, public DNA-microarray database a realistic goal?

Alvis Brazma, Alan Robinson, Graham Cameron and Michael Ashburner

Of the techniques that are being used to obtain the massive data sets of the molecules of life, the most visible is the DNA sequencing of the human genome. Following on from the publication of the human chromosome 22 sequence[1], a rough draft of the whole human genome should be available by the spring. But such advances can create the false impression that everything about life at the molecular level will soon be understood.

In reality, genome projects simply transfer digital information from DNA to computer file; this genetic 'parts-list' is a long way from providing an understanding of function. It took hundreds of years to advance from a fairly detailed understanding of human anatomy to any real understanding of function. Knowing the genome sequence and even the location of all an organism's genes is the 'anatomical' description of its genome. Functional genomics is the science of understanding how the genome functions through controlling the expression of genes. This scientific discipline will continue long after the experiment looking at 40,000 genes from 10 different samples, under 20 different conditions, produces at least 8,000,000 pieces of information. Currently, these data are scattered among various independent Internet sites, or may not be publicly available at all, although conclusions drawn from the data will have been published. Details about how experiments were carried out are often incomplete. Yet the amount of information being produced in this way is set to explode as the cost of microarray technology falls.

The need for a public repository

It is time to create a public repository for microarray data, with standardized annotation (see Box 2, overleaf). But this is a complex and ambitious project, and is one of the biggest challenges that bioinformatics has yet faced. Major difficulties stem from the detail required to describe the conditions of an experiment, and the relative and imprecise nature of measurements of expression levels. The potentially huge volume of data only adds to these difficulties. However, it is this very complexity that makes an organized repository necessary.

Important tasks to be undertaken in-

One difficulty concerns the inherent fuzziness of gene-expression data. Essentially all current expression measurements are relative: we can tell which genes are expressed differently in an experiment only in comparison with another experiment, or in relation to another gene in the same experiment. Such methods tell us little about how many copies of a messenger RNA are present. Moreover, the transcription levels reported are an average over the whole cell population sampled.

Consequently, gene-expression measurements from different technologies, or even from the same technology but from different laboratories, may not be quantitatively comparable. Two steps should allow data from different sources to be compared. First, relatively raw data should be stored to obviate any variation owing to, say, data-normalization methods. Second, standard sets of control probes and samples should be designed and used in experiments to give reference points so that these data can be normalized, at least from the same experimental platform.

The ability to compare results obtained using different technologies will depend on careful comparison of the technologies; such experiments should be encouraged and

Open access, freely available online

Community Page

Submission of Microarray Data to Public Repositories

Catherine A. Ball, Alvis Brazma*, Helen Causton, Steve Chervitz, Ron Edgar, Pascal Hingamp, John C. Matese, Helen Parkinson, John Quackenbush, Martin Ringwald, Susanna-Assunta Sansone, Gavin Sherlock, Paul Spellman, Chris Stoeckert, Yoshio Tateno, Ronald Taylor, Joseph White, Neil Winegarden

A fundamental principle guiding the publication of scientific results is that the data supporting any scholarly work must be made fully available to the research community, in a form that allows the basic conclusions to be evaluated independently. In the context of molecular biology, this has typically meant that authors of a paper describing a newly sequenced genome, gene, or protein must deposit the primary data in a permanent, public data repository, such as the sequence databases maintained by the DNA Data Bank of Japan (DDBJ), European Bioinformatics Institute (EBI), and National Center for Biotechnology Information (NCBI). Similarly, we, members of the Microarray Gene Expression Data Society (MGED; http://www.mged.org), believe that all scholarly scientific journals should now require the submission of microarray data to public repositories as part of the process of publication. While some journals have already made this a condition of acceptance, we feel that submission requirements should be applied consistently and that journals should recognize ArrayExpress (Brazma et. al. 2003), Gene Expression Omnibus (GEO) (Edgar et. al 2002), and the Center for Information Biology Gene Expression Database (CIBEX) checklist.html) as a means of achieving this goal. (2) Scientific journals should require that all primary microarray data are submitted to one of the public repositories—ArrayExpress, GEO, or CIBEX—in a format that complies with the MIAME guidelines. (3) Public databases should work with authors and scientific journals to establish data submission and release protocols to assure compliance with MIAME guidelines. (4) To assist with the review process, the databases should continue to work in collaboration with publishers to provide qualified referees with secure means of accessing prepublication data. Authors should be strongly encouraged to submit data to the databases during review.

Naturally, data should be protected from general release prior to either publication or authorization from the data submitters, whichever comes first. At a minimum, journals should require valid accession numbers for microarray data as a requirement for publication, and these accession numbers should be included in the text of the manuscript to allow members of the community to find and access the underlying data.

Since its inception in 1999, MGED has been working with the broader scientific community to establish standards for the exchange and annotation of microarray data. In

Abbreviations: CIBEX, Center for Information Biology Gene Expression Database; DDBJ, DNA Data Bank of Japan; EBI, European Bioinformatics Institute; GEO, Gene Expression Omnibus; MIAME, Minimum Information about a Microarray Experiment; MGED, Microarray Gene Expression Data Society; NCBI, National Center for Biotechnology Information

Catherine A. Ball is in the Department of Biochemistry, Stanford University School of Medicine, Stanford, California, United States of America; Alvis Brazma is at the European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, United Kingdom; Helen Causton is at the Clinical Sciences Centre/Imperial College Microarray Centre, Imperial College, London, United Kingdom; Steve Chervitz is Bioinformatics Engineer at CIS Enterprise Data Group, Affymetrix, Emeryville, California, United States of America; Ron Edgar is at the National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland, United States of America; Pascal Hingamp is at the Laboratoire Technologies Avancées pour le Génome et la Clinique, Faculte des Sciences de Luminy, Centre d'Immunologie de Marseille Luminy, Universite Aix-Marseille-II, Marseille Cedex, France; John C. Matese is at the Lewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, Princeton, New Jersey, United States of America; Helen Parkinson is at the European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, United Kingdom; John Quackenbush is at The Institute for Genomic Research, Rockville, Maryland, United States of America; Martin Ringwald is at the Jackson Laboratory, Bar Harbor, Maine, United States of America; Susanna-Assunta Sansone is at the European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, United Kingdom; Gavin Sherlock is in the Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America; Paul Spellman is Computational Scientist at Lawrence Berkeley National Laboratory, Berkeley, California,

nature

26 September 2002 Volume 419 Issue no 6905

Microarray standards at last

Not a moment too soon, the microarray community has issued guidelines that will make their data much more useful and accessible. *Nature* and the Nature research journals will respond accordingly.

You read a paper with a fascinating conclusion about the expression of several genes. You decide to use some of the same experiments on your system of choice. But when you wade through hundreds of pages of supplementary information, you find that crucial details needed for replication are missing.

Welcome to the exciting but frustrating world of DNA microarray research. Microarrays are plastic or glass chips spotted with tiny amounts of thousands of probes, used to query the activity levels of that many genes in any tissue or organism at one time. Variables in every step of the experiment often make cross-paper comparison virtually impossible. Microarray papers also pose a considerable strain on the refereeing process; the vast amounts of data mean that critical review is a monumental task.

Yet referees sometimes feel they are not given enough details, leading cautious reviewers to think that they must reanalyse the primary data set. In other cases, the primary data provided are in proprietary software and so are impossible to comment on. Many journals allowed authors to put the huge data files on their own websites for the review process, until it became clear that unscrupulous authors compromised the anonymity of referees by tracking who had visited the website.

In a move to remedy these problems, the international Microarray Gene Expression Data (MGED) group has written an open letter to scientific journals proposing standards for publication. Other members of the microarray community welcomed these steps, designed to clarify the Minimal Information About a Microarray Experiment (MIAME) guidelines (*Nature Genetics* **29**, 365–371; 2001).

For authors, the proposal provides a checklist of variables that should be included in every microarray publication, at http://www.mged.org/Workgroups/MIAME/miame_checklist.html. This checklist, with all variables completed, would be supplied as supplementary information at the time of submission. The MGED group suggests that journals require submission of microarray data to either of two databases emerging as the main public repositories: GEO (www.ncbi.nlm.nih.gov/geo/) or ArrayExpress (www.ebi.ac.uk/arrayexpress).

Harried editors can rejoice that, at last, the community is taming the unruly beast that is microarray information. Therefore, all submissions to *Nature* and the Nature family of journals received on or after 1 December containing new microarray experiments must include the mailing of five compact disks to the editor. These disks should include necessary information compliant with the MIAME standard. The information must be supplied in a format that could be read by widely available software packages. Data integral to the paper's conclusions should be submitted to the ArrayExpress or GEO databases, with accession numbers where available, supplied at or before acceptance for publication.

How much data should authors provide to the community? Specifically, do other researchers really need to recreate the exact microarray just to test the expression level of a few key genes, which could presumably be done through other methods? Perhaps with further evolution and standardization of microarray technology, the need to specify so many variables will decrease, but the MGED standards are surely appropriate for the current state of the field. ■

ARRAYEXPRESS

The **ArrayExpress Archive** is a database of functional genomics experiments including gene expression where you can query and download data collected to MIAME and MINSEQE standards. **Gene Expression Atlas** contains a subset of curated and re-annotated Archive data which can be queried for individual gene expression under different biological conditions across experiments.

Experiments Archive

10043 experiments, 280785 assays

Experiment, citation, sample and factor annotations

Query

Browse experiments
Advanced query interface

Submitter/reviewer login

ArrayExpress Query Help

Gene Expression Atlas

1236 experiments, 33681 assays, 6446 conditions

Genes | up/down in | Conditions

Any species

Query

Gene Expression Atlas Home

News

- *21 Dec 2009* - **ArrayExpress Bioconductor package**

  The latest version of the ArrayExpress Bioconductor package is required to access the data file archives as the naming convention used has changed. There is more information in the new FAQ page.

Links

- ArrayExpress **User Survey**
- **Help** | **Training** | **FAQ** | **Citing**
- **Submit Data** (array based and re-sequencing)
- **Programmatic Access** | **FTP Access**
- **Software Downloads** and **Statistics**
- **EFO** | **Bioconductor Package** | **Quality Metrics**

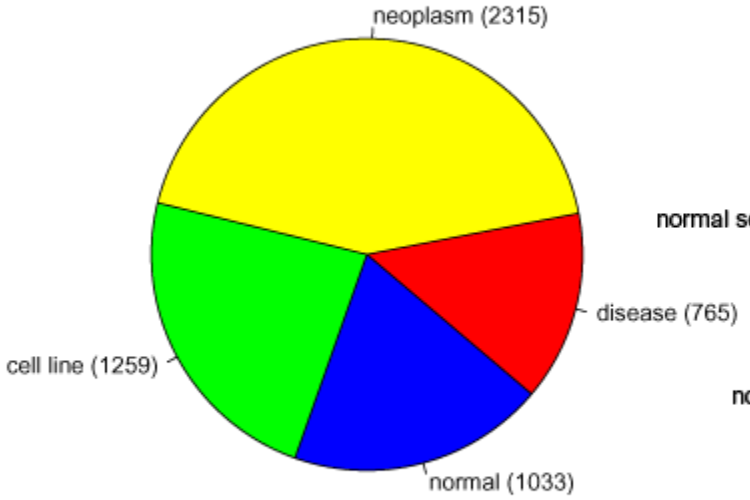
Enabling research that would not be possible otherwise

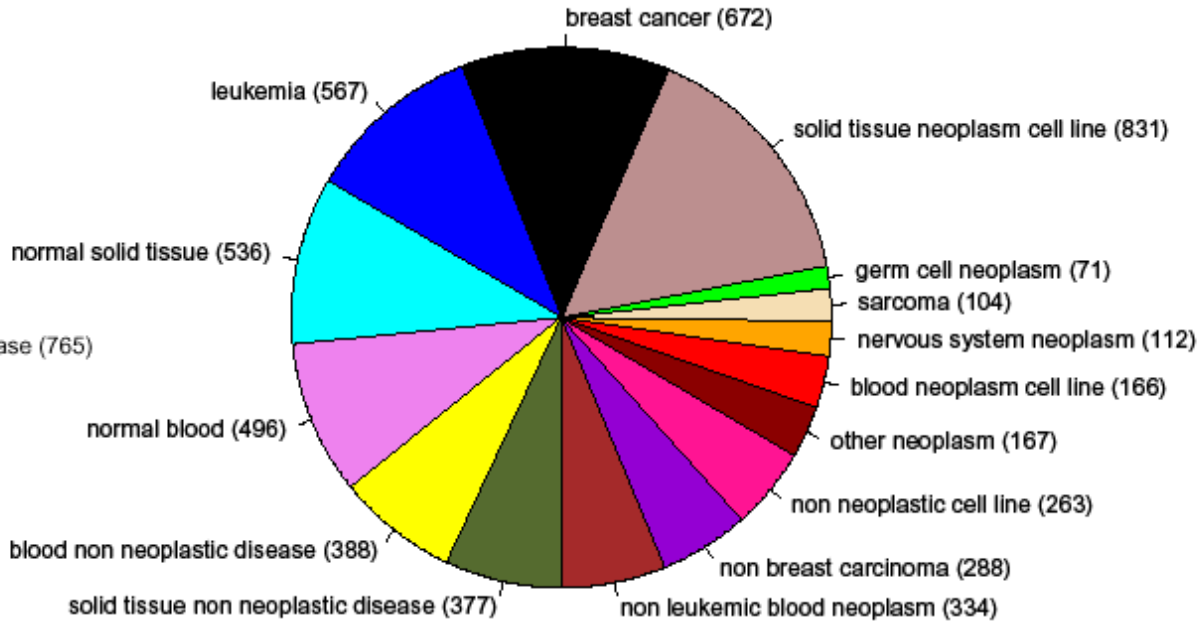
- We collected over 9000 raw data files from Affymetrix U133A from GEO and ArrayExpress
- Applying strict quality controls, removing the duplicates
- Data on **5372** samples remained

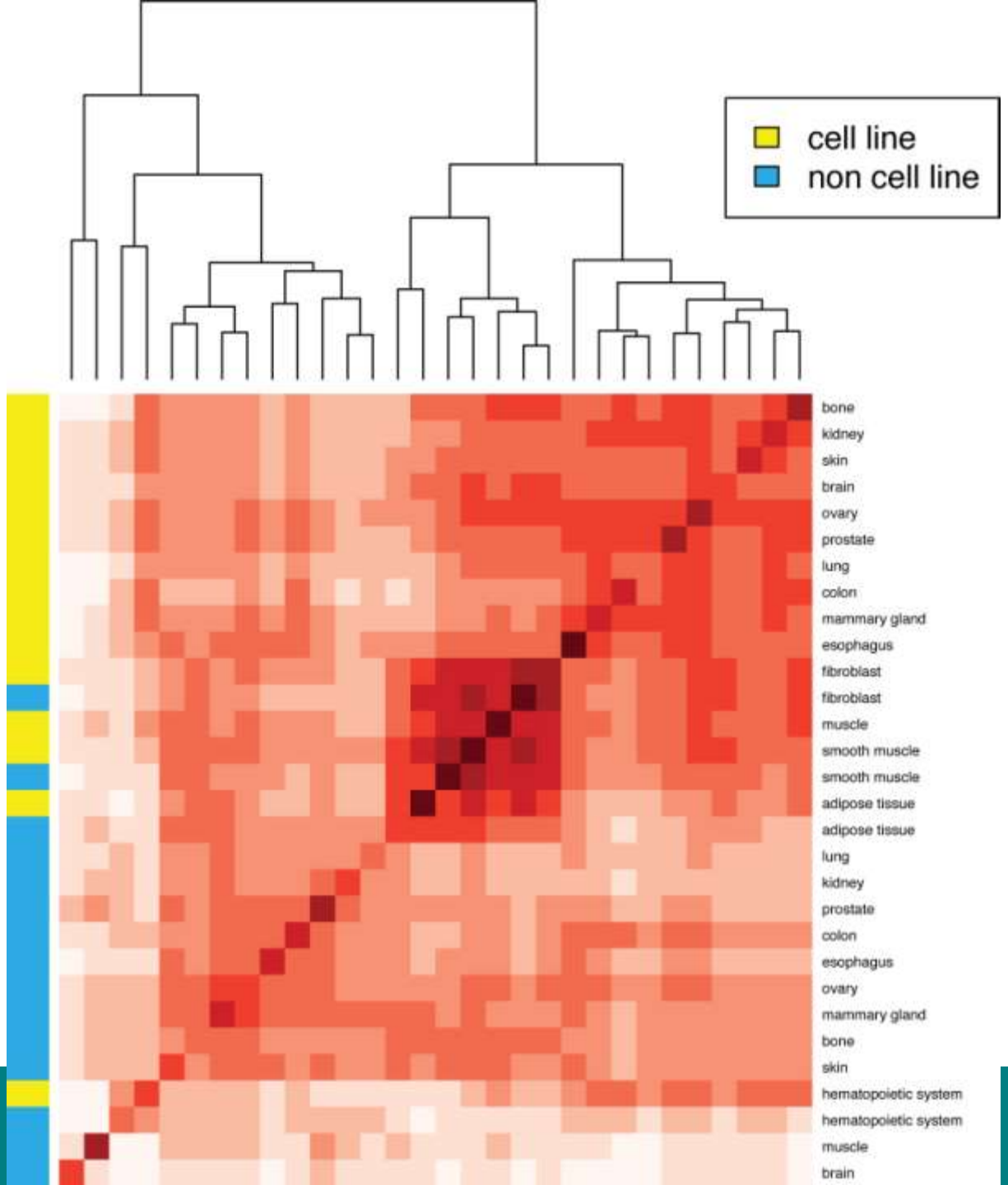
  from **206** different studies generated

  in **163** different laboratories

  grouped in **369** different biological ‘conditions’ (tissue types, diseases, various cell lines, etc)

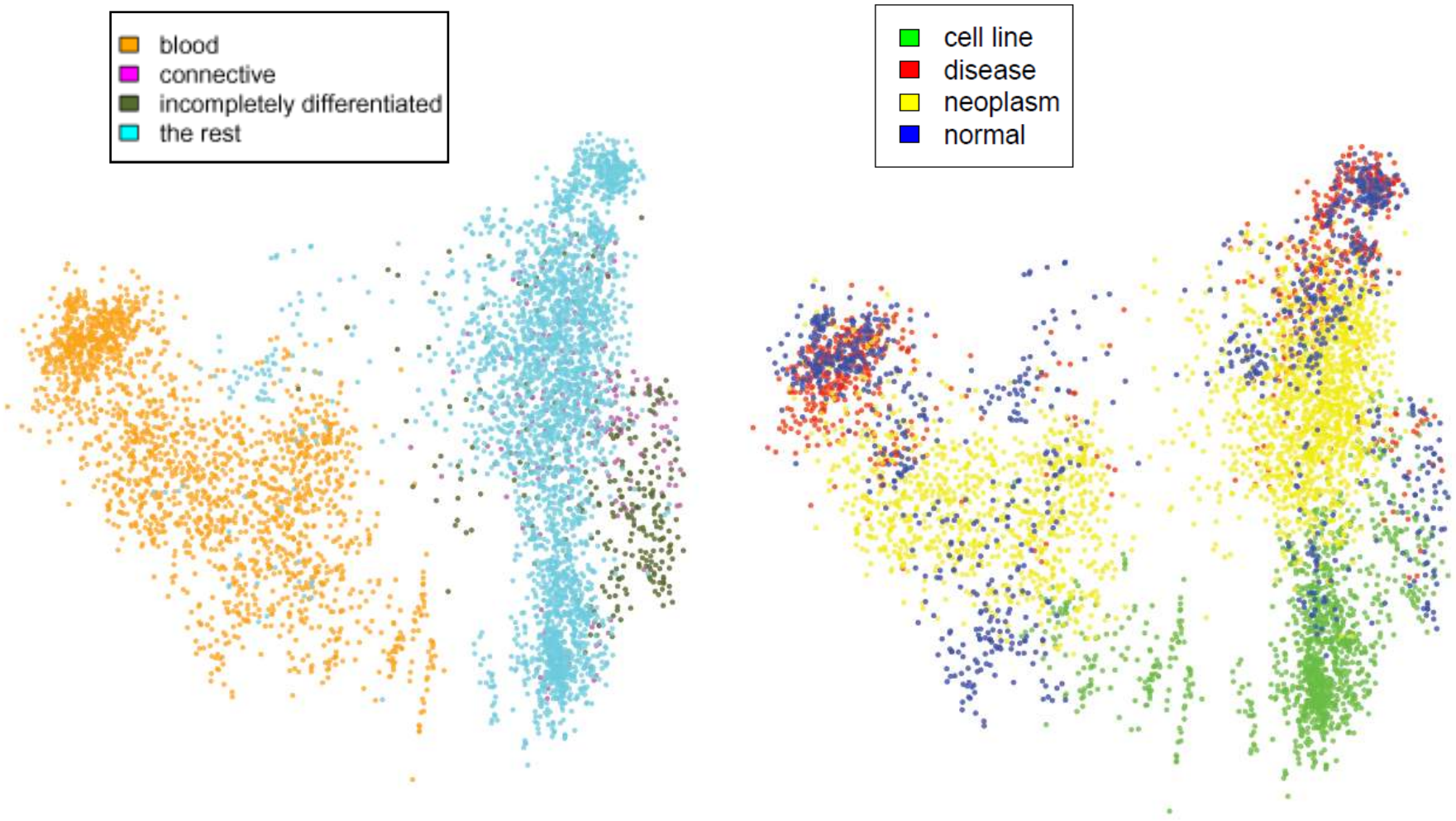

neoplasm (2315)
disease (765)
cell line (1259)
normal (1033)


breast cancer (672)
leukemia (567)
solid tissue neoplasm cell line (831)
normal solid tissue (536)
germ cell neoplasm (71)
sarcoma (104)
nervous system neoplasm (112)
blood neoplasm cell line (166)
normal blood (496)
other neoplasm (167)
non neoplastic cell line (263)
blood non neoplastic disease (388)
non breast carcinoma (288)
solid tissue non neoplastic disease (377)
non leukemic blood neoplasm (334)



Gene expression wise most cell lines are very similar and rather different from their tissues or origin!

Hematopoietic and malignancy axes



Lukk et al, Nature Biotechnology, 28: 322

What did MIAME achieve

- To guide microarray software development – all microarray data shares a certain structure
- To help to implement the principle that data supporting publications must be made available to the public in a usable format
- Enabled new conclusions drawn from metaanalysis of combined data

Being modest in goals and agile in implementation was the key!

MGED standards – MAGE-ML

Research

Design and implementation of microarray gene expression markup language (MAGE-ML)

Paul T Spellman[1], Michael Miller[2], Jason Stewart[3], Charles Troup[4], Ugis Sarkans[5], Steve Chervitz[6], Derek Bernhart[6], Gavin Sherlock[7], Catherine Ball[7], Marc Lepage[8], Marcin Swiatek[9], WL Marks[10], Jason Goncalves[10], Scott Markel[11], Daniel Iordan[10], Mohammadreza Shojatalab[5], Angel Pizarro[12], Joe White[13], Robert Hubley[14], Eric Deutsch[14], Martin Senger[5], Bruce J Aronow[15], Alan Robinson[5], Doug Bassett[2], Christian J Stoeckert Jr[12] and Alvis Brazma[5]

It didn't work!

Addresses: [1]Department of Cell and Molecular Biology, University of California at Berkeley, Berkeley, CA 94720-3206, USA. [2]Rosetta Biosoftware, 113th Ave NE, Kirkland, WA 98034, USA. [3]Open Informatics, Arizona St SE, Albuquerque, NM 87108, USA. [4]Bioscience Research - Agilent Technologies, Deer Creek Rd, Palo Alto, CA 94304, USA. [5]European Bioinformatics Institute, EMBL Hinxton Outstation, Cambridge CB10 1SD, UK. [6]Affymetrix, Inc., Vallejo St, Emeryville, CA 94608, USA. [7]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305-5120, USA. [8]Molecular Mining Corporation, Rideau St, Kingston, ON K7K 2Z8, Canada. [9]Imaging Research Inc., Glenridge Ave, St. Catharines, ON L2S 3A1, Canada. [10]Iobion Informatics LLC, North Torrey Pines Road, La Jolla, CA 92037, USA. [11]LION bioscience Inc., Executive Drive, San Diego, CA 92121, USA. [12]Center for Bioinformatics, University of Pennsylvania, Guardian Drive, Philadelphia, PA 19104, USA. [13]The Institute for Genomic Research, Medical Center Drive, Rockville, MD 20850, USA. [14]Computational Biology, Institute for Systems Biology, North 34th St, Seattle, WA 98103-8904, USA. [15]CHRF, Burnet Ave, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA.

Correspondence: Paul T Spellman. E-mail: spellman@fruitfly.org

BMC Bioinformatics

BioMed Central

Commentary

Open Access

A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB

Tim F Rayner*[1], Philippe Rocca-Serra[1], Paul T Spellman[2], Helen C Causton[3], Anna Farne[1], Ele Holloway[1], Rafael A Irizarry[4], Junmin Liu[5], Donald S Maier[6], Michael Miller[7], Kjell Petersen[8], John Quackenbush[9], Gavin Sherlock[10], Christian J Stoeckert Jr[5], Joseph White[9], Patricia L Whetzel[5], Farrell Wymore[6], Helen Parkinson[1], Ugis Sarkans[1], Catherine A Ball[6] and Alvis Brazma*[1]

Address: [1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [2]Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [3]MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College Hammersmith Hospital Campus, London, UK, [4]Department of Biostatistics, John Hopkins School of Public Health, Baltimore, MD, USA, [5]Center for Bioinformatics and Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA, USA, [6]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, USA, [7]Rosetta Biosoftware, Rosetta Inpharmatics, LLC, Seattle, USA, [8]Bergen Center for Computational Science, Computational Biology Unit, University of Bergen, Norway, [9]Dana Farber Cancer Institute, Boston, MA, USA and [10]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

Email: Tim F Rayner* - rayner@ebi.ac.uk; Philippe Rocca-Serra - rocca@ebi.ac.uk; Paul T Spellman - ptspellman@lbl.gov;

It did work – used in ArrayExpress, Bioconductor, caArray, …

A lesson

- Things should be made as simple as possible, but no simpler (A. Einstein)

Proliferation of functional genomics standards

- MIAME
- MIAPE
- MIACA
- MIARE
- MISFISIE
- …

- Standardisation of standards - MIBBI

COMMENTARY

Nature Biotechnology volume 26 number 8 AUGUST 2008

ublishing Group http://www.nature.com/naturebiotechnology

Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project

Chris F Taylor[*1,2], Dawn Field[2,3], Susanna-Assunta Sansone[1,2], Jan Aerts[4], Rolf Apweiler[1], Michael Ashburner[5], Catherine A Ball[6], Pierre-Alain Binz[7,8], Molly Bogue[9], Tim Booth[2], Alvis Brazma[1], Ryan R Brinkman[10], Adam Michael Clark[11], Eric W Deutsch[12], Oliver Fiehn[13], Jennifer Fostel[14], Peter Ghazal[15], Frank Gibson[16], Tanya Gray[2,3], Graeme Grimes[15], John M Hancock[17], Nigel W Hardy[18], Henning Hermjakob[1], Randall K Julian Jr[19], Matthew Kane[20], Carsten Kettner[21], Christopher Kinsinger[22], Eugene Kolker[23,24], Martin Kuiper[25], Nicolas Le Novère[1], Jim Leebens-Mack[26], Suzanna E Lewis[27], Phillip Lord[16], Ann-Marie Mallon[17], Nishanth Marthandan[28], Hiroshi Masuya[29], Ruth McNally[30], Alexander Mehrle[31], Norman Morrison[2,32], Sandra Orchard[1], John Quackenbush[33], James M Reecy[34], Donald G Robertson[35], Philippe Rocca-Serra[1,36], Henry Rodriguez[22], Heiko Rosenfelder[31], Javier Santoyo-Lopez[15], Richard H Scheuermann[28], Daniel Schober[1], Barry Smith[37], Jason Snape[38], Christian J Stoeckert Jr[39], Keith Tipton[40], Peter Sterk[1], Andreas Untergasser[41], Jo Vandesompele[42] & Stefan Wiemann[31]

More lessons for standards in basic research

- You can’t coerce scientists (or anyone) to use standards
- Assess realistically what the implementation is going to cost and how much time it will take
- Most successful standards are developed bottom up
- Successful standards are often defined by killer software applications (bottom up standards)
- Approval of community, journals and funding agencies
- Keep the standards simple
- It’s moving target, life cycle may be short, be ready to redevelop everything soon
- Open standards (and open source software)

SCIENTIFIC CONFERENCES
HINXTON CAMBRIDGE UK

wellcome trust

Bridging the Gap in Biomedical Genetics

27–29 October 2010

Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

BACKGROUND

Biomedical research is increasingly using new molecular technologies, such as genotyping, transcription profiling, or metabolite profiling to understand disease mechanisms, test drug efficacy, and develop new therapies. Patient samples are now routinely profiled by such technologies and the data are collected in biomedical research warehouses for analysis. The patient health records are typically managed in separate information systems. To achieve the full power of the new approaches for translational research, the molecular data need to be integrated with information from the patient records.

SCIENTIFIC ORGANISING COMMITTEE:

Janet Thornton European Bioinformatics Institute, UK
Alvis Brazma European Bioinformatics Institute, UK
David Ingram University College London, UK
Johan Van der Lei Erasmus University Medical Center, Netherlands

KEYNOTE SPEAKERS:

Norbert Graf University of Saarland, Germany
Isaac Kohane Harvard MIT, USA
Bernadette Modell University College London, UK
Gertjan van Ommen Leiden University Medical Center, Netherlands

Acknowledgements

- Margus Lukk
- **Misha Kapushesky**
- Angela Gonzales
- **Helen Parkinson**
- Gabriela Rustici
- **Ugis Sarkans**
- Ele Holloway
- Roby Mani
- Mohammadreza Shojatalab
- Nikolay Kolesnikov
- Niran Abeygunawardena
- Anjan Sharma
- Miroslaw Dylag
- Ekaterina Pilicheva
- Ibrahim Emam
- Pavel Kurnosov
- Andrew Tikhonov
- Andrey Zorin
- Anna Farne
- Eleanor Williams
- Tony Burdett
- James Malone
- Holly Zheng
- Tomasz Adamusiak
- Susanna-Assunta Sansone
- Philippe Rocca-Serra
- Natalija Sklyar
- Marco Brandizi
- Chris Taylor
- Eamonn Maguire
- Maria Krestyaninova
- Mikhail Gostev
- Johan Rung
- Natalja Kurbatova
- Katherine Lawler
- Nils Gehlenborg
- Lynn French

- Collaborators
  - Audrey Kaufman (EBI)
  - Wolfgang Huber (EBI)
  - Sami Kaski (Helsinki)
  - Morris Swertz (Groningen)
  - …
- Funding
  - European Commision
    - FELICS
    - MolPAGE
    - ENGAGE
    - MuGEN
    - SLING
    - DIAMONDS
    - EMERALD
  - NIH (NHGRI)
  - EMBL